Erschienen in: Storrer, Angelika/Geyken, Alexander/Siebert, Alexander/Würzner, Kay-Michael (Hrsg.): KONVENS 2008 – Ergänzungsband. Textressourcen und lexikalisches Wissen. - Berlin: BBAW, 2008. S. 21-27.

Demonstration des SemDok-Textparsers

Mirco Hilbert, Harald Lüngen, Maja Bärenfänger, Henning Lobin

Zusammenfassung. Im Teilprojekt C1 "SemDok" der DFG-Forschergruppe *Texttechnologische Informationsmodellierung* wurde ein Textparser für Diskursstrukturen wissenschaftlicher Zeitschriftenartikel nach der Rhetorical Structure Theory entwickelt. Die wesentlichen konzeptuellen und technischen Merkmale des Chart-Parsers und die sich daraus ergebenden Parametrisierungsmöglichkeiten für Parsing-Experimente werden beschrieben. Zudem wird *HPViz*, ein Tool für die Visualisierung von Parsing-Ergebnissen (RST-Bäume in einer XML-Anwendung) und die Navigation in ihnen, vorgestellt.

1 Einleitung

Im Teilprojekt C1 "SemDok" der DFG-Forschergruppe 437 *Texttechnologische Informationsmodellierung* wurde ein Textparser für Diskursstrukturen wissenschaftlicher Zeitschriftenartikel nach der *Rhetorical Structure Theory* (RST, Mann und Thompson 1988; Mann und Taboada 2005) entwickelt. Eine Dokumentation der Konzeption, Prolog-Implementierung und Evaluation dieses Systems findet sich in der bevorstehenden Buchpublikation Mehler et al. (2009). Der vorliegende Beitrag beinhaltet eine Beschreibung von Features und Funktionalitäten des Parsers, die für Evaluationsläufe parametrisiert werden können, sowie des Visualisierungs- und Navigationstools *HPViz* für RST-Bäume, d.h. Parsingergebnisse.

2 SemDok-Textparser

2.1 Szenario

Bei der Entwicklung des SemDok-Parsers wurde das Szenario einer interaktiven Lernkomponente zu Grunde gelegt, mit der Studierende selektive und explorative Lesestrategien für wissenschaftliche Papiere einüben können. Sie werden dabei von dem System durch zwei wesentliche Mechanismen unterstützt: 1.) automatisches Highlighten von Textstrukturen und rhetorisch oder thematisch salienten Textsegmenten; 2.) Bereitstellung automatisch generierter Hyperlinks zur Navigation zwischen Textsegmenten, die in einer Kohärenzbeziehung stehen, inklusive Link-Listen zu relevanten Einstiegspunkten (Hypertextualisierung). Um auch eigene Texte (z.B.

ausgewählte Artikel, Seminararbeiten) hochzuladen und in ihnen navigieren zu können, muss das System den Texten automatisch eine Annotation ihrer Diskursstruktur hinzufügen; eine solche Funktionalität wird durch den SemDok-Parser realisiert. Die in Abschnitt 3 vorgestellte Visualisierungskomponente für RST-HP-Dokumente (Output-Format des SemDok-Parsers) orientiert sich an diesem Szenario (Bärenfänger et al. 2006; Lüngen et al. 2006a,b).

2.2 Features des SemDok-Parsers

Für die technische Realisation wurde eine Architektur gewählt, in der Analysen ein und desselben Texts auf verschiedenen linguistischen Ebenen von Präprozessoren oder projekt-externer Software bereit gestellt werden. Neben aus bisherigen Ansätzen zum RST-Diskursparsing (vgl. Corston-Oliver 1998; Marcu 2000) bekannten Typen von Wissensquellen wie lexikalische Diskursmarker, Interpunktion, syntaktische und morphologische Merkmale, werden für das Textparsing wissenschaftlicher Zeitschriftenartikel in SemDok auch Textanalysen auf höheren linguistischen Ebenen zur Desambiguierung von Diskursrelationen einbezogen, speziell Analysen der thematischen Struktur (in Form von Anaphernresolution und lexikalischen Ketten), der logischen Dokumentstruktur und der generischen Texttypstruktur. Nach dem Konzept der XML-basierten Mehrebenen-Annotation (Witt 2004) werden die jeweiligen Analysergebnisse als XML-Annotationen repräsentiert und im Parser durch die Einbindung der sogenannten Sekimo-Tools (Witt et al. 2005) ausgewertet, indem Konfigurationen vom Elementen auf den unterschiedlichen Analyseebenen in den Bedingungen für *Reduce*-Operationen geprüft werden.

Abbildung 1 zeigt die Architektur des SemDok-Parsers. Der mittlere Bereich zeigt die sieben Analyseebenen, auf der XML-Annotationen von vorverarbeitenden Komponenten bereit gestellt werden. Einige der Komponenten verwenden bestimmte Wissensquellen, die auf der linken Seite dargestellt sind. Jede Komponente fügt dem Eingabetext eine Annotationsschicht hinzu, allein die unteren beiden Annotationsschichten "LC" und "TTS" sind in der aktuellen Version noch nicht realisiert. Der Parser liest alle Annotationsschichten sowie den Originaltext in Gestalt einer Prolog-Faktenbasis nach dem Format in Witt et al. (2005) ein. Einzig die Annotationsschicht "SEG" ist obligatorisch, sie enthält die initiale Segmentierung in elementare und komplexe Diskurssegmente und steuert die zentrale Parsingkomponente. Der Parser gibt ein RST-HP-Dokument aus. RST-HP ist das in SemDok entwickelte XML-Annotationsformat zur Repräsentation von RST-Bäumen. Der Parser fügt also vorhandenen n Annotationsschichten eine weitere Annotationsschicht, die Schicht $n+1$, hinzu.

Die zentrale Parsingkomponente heißt *GAP* (*Generalised Annotation Parser*), es handelt sich um einen in Prolog implementierten Bottom-up Passive Chart Par-

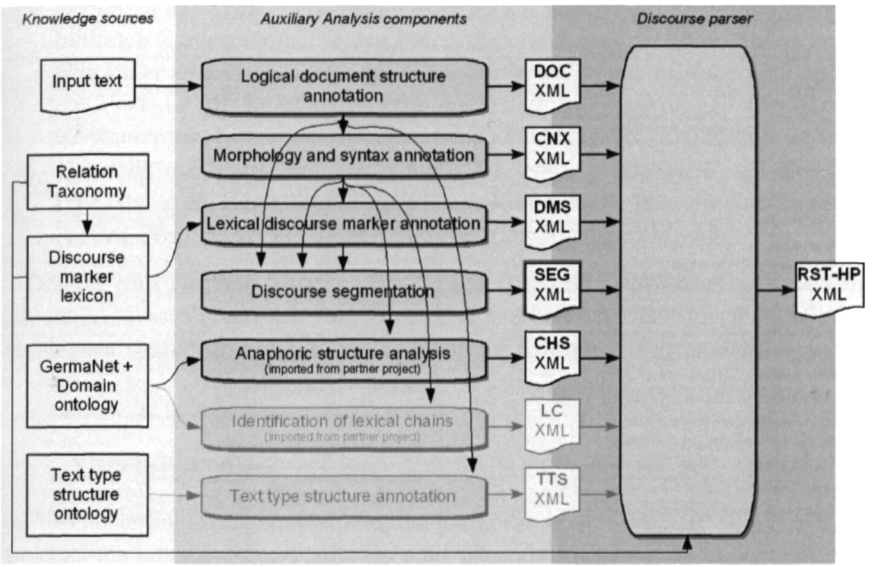

Abbildung 1. Architektur des Textparsers

ser. GAP-Aufrufe werden für jedes in der initialen Segmentierung gefundene *containing element* iteriert, eine solche Iteration wird in einer Kaskade wiederum für jedes Dokument-Level gestartet. So wird GAP zunächst auf dem Dokument-Level "EDS+" für jedes *containing element* vom Typ "SDS" (*sentential discourse segment*) aufgerufen, um Elemente vom Typ "EDS" (*elementary discourse segments*, in etwa: Teilsätze) zu kombinieren. In der zweiten Kaskadenstufe "SDS+" werden SDS zu CDS_{block} (*complex discourse segment* vom Typ *block*, in etwa: Paragraphen-Level) kombiniert, in der dritten CDS_{block} zu CDS_{div} und schließlich CDS_{div} innerhalb des top-level Segmenttyps CDS_{doc}. Die *containing element*s stellen somit Top-down-Constraints im ansonsten bottom-up verlaufenden Textparsing dar.

Für Parsing-Experimente können insgesamt folgende Parametrisierungen des Textparsers vorgenommen werden:

1. *Auswahl eines Relations-Sets.* Zur Auswahl steht ein RST-Relations-Set für die Analyse wissenschaftlicher Zeitschriftenartikel mit 44 Kategorien sowie ein reduziertes Set mit 30 Kategorien.
2. *Zuwahl oder Abwahl von Annotationsschichten.* Wird eine Annotationsschicht abgewählt, so werden diejenigen Reduce-Regeln, die sich auf sie beziehen, nicht angewendet. Diese Parameter dienen der Evaluation des Beitrags von

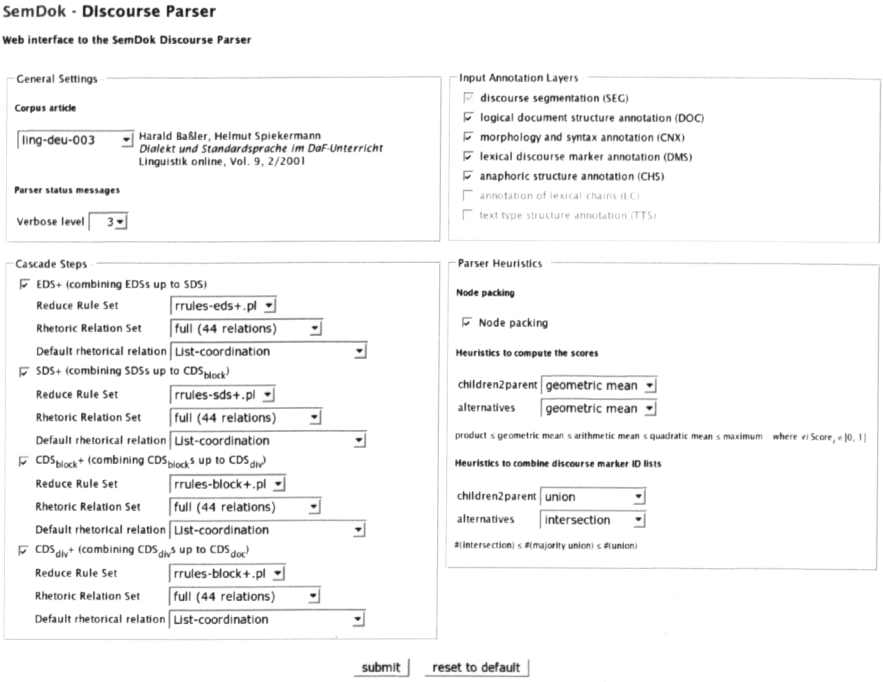

Abbildung 2. Web-Interface

bestimmten Annotationsschichten zum Gesamtergebnis.

3. Kaskadenstufen-relevante Parameter, die vom Benutzer in sinnvoller Weise kombiniert werden müssen:

   (a) *Zuwahl oder Abwahl von Kaskadenstufen.* Wird eine Kaskadenstufe abgewählt, so wird für die betreffenden Input-Segmente lediglich ein trivialer RST-Baum erstellt.

   (b) *Auswahl eines Regel-Files je Dokument-Level (Kaskadenstufe).*

   (c) *Spezifikation einer Default-Relation je Kaskadenstufe.* Bei Zuwahl wird jedem getesteten Segment-Paar, das nicht durch eine Reduce-Regel zu einem übergeordneten Segment kombiniert werden kann, eine Default-Diskursrelation zugeordnet. Zur Auswahl stehen die im Entwicklungskorpus am häufigsten anzutreffenden rhetorischen Relationen. Diese Parameter dienen der Ermittlung von Baselines für Evaluationen.

4. *Zuwahl oder Abwahl des Packens von Kanten.*

5. *Verfahren der Score-Berechnung* einer neuen Kante aus den Scores der kombinierten Kanten für die beiden Fälle "children2parent" (Kombination zweier adjazenter Diskurssegmente zu einem größeren Segment (Magerman & Marcus 1991; Le Thanh & Abeysinghe 2003) und "alternatives" (Eintrag gepackter Kanten zur Kombination konkurrierender Analysen, (vgl. Tomita 1987).
6. *Heuristik zur Ermittlung der DM-ID-Liste einer gepackten Kante.* Die DM-ID-Liste ist die Komponente einer Chart-Kante, in der über die im Verlauf der Parsing-Historie verwendeten Diskursmarker Buch geführt wird.

Derzeit wird ein Web-Interface für die Steuerung des SemDok-Textparsers implementiert, dessen Oberfläche im Screenshot in Abbildung 2 zu sehen ist. Die vorgeschlagene Demonstration des Systems auf der Konvens erfolgt mittels dieses Interfaces.

3 Visualisierung

Die Visualisierungskomponente dient der Exploration von RST-HP-Strukturen, die das Ergebnis eines Parsing-Prozesses sind oder als Master-Annotationen manuell angefertigt wurden. Sie realisiert somit in prototypischer Form einen Teil des in Abschnitt 2.1 beschriebenen Szenarios. In einer geteilten Ansicht werden auf der linken Seite die RST-HP-Baumstruktur eines Textsegments sowie auf der rechten Seite der zugehörige Text dargestellt.

Im Gegensatz zu traditionellen RST-Bäumen, wie sie mit dem RSTTool von Mick O'Donnell konstruiert und visualisiert werden können (O'Donnell 2000), wird die RST-HP-Struktur als *discourse dependency tree* im Sinne von Danlos (2005) dargestellt, wobei rhetorische Relationen als Knoten und die Nuklearität als Kantenbeschriftung ausgedrückt werden.

Der Benutzer hat nun mehrere Möglichkeiten, die RST-HP-Struktur des Textes zu explorieren. Auf Seiten des RST-HP-Baums ist jeder Knoten und jede Kante anklickbar, so dass der Benutzer die zugehörige Textstelle anspringen kann. Nuklei einer ausgewählten Relation werden dann im Text rot, Satelliten grün hervorgehoben. In der Textansicht können Sätze, Paragraphen oder ganze Abschnitte ausgewählt werden, um deren RST-HP-Strukturen anzeigen zu lassen. Sätze werden beim Überfahren mit der Maus hervorgehoben und können so angeklickt werden. Die RST-Struktur eines ausgewählten Satzes wird dann im Kontext seines Paragraphen dargestellt, sofern der zugehörige Relationsknoten nicht bereits Teil des angezeigten Baums ist.

Die hierarchische Textstruktur bestehend aus Abschnitten, Paragraphen und weiteren Block-Elementen wird seitlich des Textes durch Einrückung mittels senkrechter Balken dargestellt. Auch diese Balken sind maus-sensitiv und können angeklickt werden, um die RST-HP-Struktur eines kompletten Paragraphen anzeigen zu lassen.

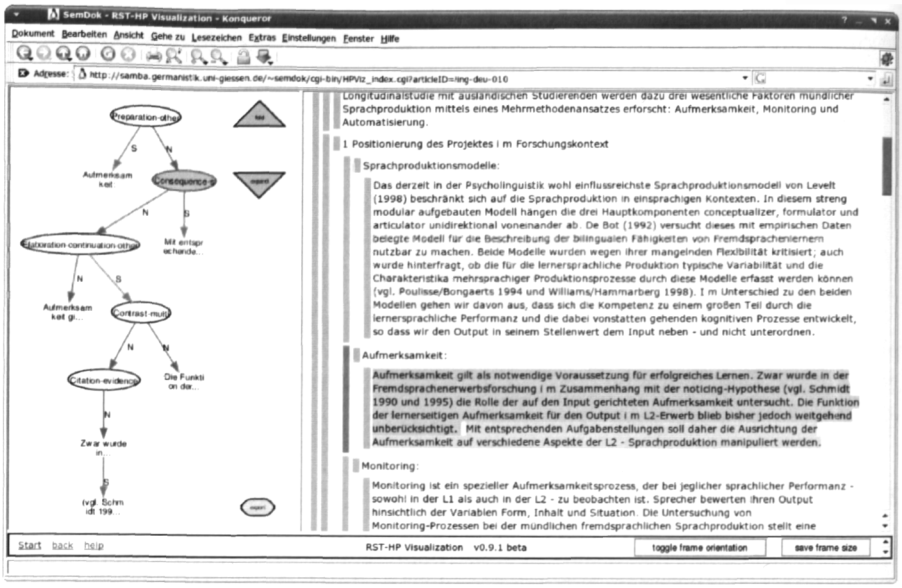

Abbildung 3. RST-HP Visualisierung

Die RST-HP Visualisierung HPViz ist ein Web-Interface, das mit CGI/Perl umgesetzt wurde, und benutzt XSLT und Prologkomponenten zur Vorbereitung der Eingabedaten, XSLT und WebDot/Dot zur Umwandlung und Darstellung des RST-HP-Baums und des Textes sowie Image-Maps, JavaScript und CSS zur Realisierung der gegenseitigen Verlinkung und Navigation.

Literaturverzeichnis

Bärenfänger, Maja, Mirco Hilbert, Henning Lobin, Harald Lüngen, und Csilla Puskàs (2006). Cues and constraints for the relational discourse analysis of complex text types - the role of logical and generic document structure. In: Candy Sidner, John Harpur, Anton Benz, Peter Kühnlein, und Candy Sidner (Hrsg.), *Proceedings of the Workshop on Constraints in Discourse*, 27–34, Maynooth, Ireland: National University of Ireland.

Corston-Oliver, Simon (1998). *Computing of Representations of the Structure of Written Discourse*. Doktorarbeit, University of California, Santa Barbara.

Danlos, Laurence (2005). Comparing RST and SDRT discourse structures through dependency graphs. In: Claudia Sassen, Anton Benz, und Peter Kühnlein (Hrsg.), *Proceedings of Constraints in Discourse*, 55–62, Dortmund.

Le Thanh, Huong und Geetha Abeysinghe (2003). A study to improve the efficiency of a discourse parsing system. In: A. Gelbukh (Hrsg.), *Proceedings of the 4th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing'03)*, Band 2588 , *Lecture Notes in Computer Science (LNCS)*, 104–117, Mexico City, Mexico: Springer-Verlag.

Lüngen, Harald, Henning Lobin, Maja Bärenfänger, Mirco Hilbert, und Csilla Puskàs (2006a). Text parsing of a complex genre. In: *Proceedings of the Conference on Electronic Publishing (ELPUB)*, 247–256, Bansko, Bulgaria.

Lüngen, Harald, Csilla Puskàs, Maja Bärenfänger, Mirco Hilbert, und Henning Lobin (2006b). Discourse segmentation of German written text. In: *Proceedings of the 5th International Conference on Natural Language Processing (FinTAL 2006)*, 245–256, Åbo, Finland: Springer.

Magerman, David M. und Mitchell P. Marcus (1991). Pearl: A probabilistic chart parser. In: *Proceedings of the European ACL Conference*, 40–47.

Mann, William C. und Maite Taboada (2005). RST – Rhetorical Structure Theory. W3C page, `http://www.sfu.ca/rst`.

Mann, William C. und Sandra A. Thompson (1988). Rhetorical Structure Theory: Toward a functional theory of text organisation. *Text* 8(3):243–281.

Marcu, Daniel (2000). *The Theory and Practice of Discourse Parsing and Summarization*. Cambridge, MA: MIT Press.

Mehler, Alexander, Kai-Uwe Kühnberger, Harald Lüngen, Angelika Storrer, und Andreas Witt (Hrsg.) (2009). *Modelling, Processing, and Learning of Text-Technological Data Structures.* , in Vorbereitung.

O'Donnell, Michael (2000). RSTTool 2.4 – A markup tool for Rhetorical Structure Theory. In: *Proceedings of the International Natural Language Generation Conference (INLG'2000)*, 253 – 256, Mitzpe Ramon, Israel.

Tomita, Masaru (1987). An efficient augmented-context-free parsing algorithm. *Computational Linguistics* 13(1-2):31–46.

Witt, Andreas (2004). Multiple hierarchies: New aspects of an old solution. In: *Proceedings of the Extreme Markup Languages*, Montreal.

Witt, Andreas, Harald Lüngen, Daniela Goecke, und Felix Sasaki (2005). Unification of XML documents with concurrent markup. *Literary and Linguistic Computing* 20(1):103–116.